PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 109–120

Is Neural Machine Translation the New State of the Art?

Sheila Castilho,[a] Joss Moorkens,[a] Federico Gaspari,[a] Iacer Calixto,[a]
John Tinsley,[b] Andy Way[a]

[a] ADAPT Centre, Dublin City University
[b] Iconic Translation Machines

Abstract

This paper discusses neural machine translation (NMT), a new paradigm in the MT field, comparing the quality of NMT systems with statistical MT by describing three studies using automatic and human evaluation methods. Automatic evaluation results presented for NMT are very promising, however human evaluations show mixed results. We report increases in fluency but inconsistent results for adequacy and post-editing effort. NMT undoubtedly represents a step forward for the MT field, but one that the community should be careful not to oversell.

1. Introduction

Since its inception, different theories and practices for Machine Translation (MT) have come and gone, with each new wave generating great excitement and anticipation in the field. From the first commercial rule-based systems to more recent statistical models, there has, however, generally been great discrepancy between the high expectation of what MT should accomplish and what it is actually able to deliver. More recently, the neural approach (NMT) has emerged as a new paradigm in MT systems, raising interest in academia and industry by outperforming phrase-based statistical systems (PBSMT), based largely on impressive results in automatic evaluation (Bahdanau et al., 2015; Sennrich et al., 2016; Bojar et al., 2016). But do NMT results also surpass those of SMT when using human evaluation? Can we claim at this stage that NMT is the new state-of-the-art paradigm for production? This paper discusses the quality of NMT systems when compared to the state-of-the-art SMT

 Corresponding author: sheila.castilho@adaptcentre.ie
Cite as: Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, Andy Way. Is Neural Machine Translation the New State of the Art?. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 109–120. doi: 10.1515/pralin-2017-0013.

systems, by reporting on three use cases in which human evaluators compared NMT and SMT output for a range of language pairs. Based on the findings, we argue that even though NMT shows significant improvements for some language pairs and specific domains, there is still much room for research and improvement before broad generalisations can be made.

The remainder of the paper is organised as follows: in Section 2, we survey the existing literature concerning NMT systems. In Section 3, we describe three use cases where NMT systems were compared against SMT systems and human evaluation was carried out: Section 3.1 presents a study using images to machine-translate user-generated e-commerce product listings with two NMT and one SMT systems for the English-German language pair; Section 3.2 reports a small-scale human evaluation focusing on the patent domain for the Chinese language, and Section 3.3 describes a large-scale human evaluation for the MOOC domain, considering translations from English into four target languages (German, Greek, Portuguese and Russian). Finally, in Section 4, we discuss the main findings of the use cases, zooming in on how NMT was evaluated, and we draw our main conclusions of interest to the broader MT community, including developers and users.

2. The Rise of Neural Machine Translation Models

Neural models involve building an end-to-end neural network that maps aligned bilingual texts which, given an input sentence X to be translated, is normally trained to maximise the probability of a target sequence Y without additional external linguistic information. Recently, a surge of interest in NMT came with the application of deep neural networks (DNNs) to build end-to-end *encoder–decoder* models (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014). Bahdanau et al. (2015) first introduced an attention mechanism into the NMT encoder–decoder framework which is trained to attend to the relevant source-language words as it generates each word of the target sentence. Some important recent developments in NMT involve improving the attention mechanism, including linguistic information or including more languages into the model (Luong et al., 2015; Sennrich and Haddow, 2016)

NMT improvements over PBSMT systems have been reported in shared tasks, where NMT ranked above SMT systems in six of 12 language pairs for translation tasks (Bojar et al., 2016). In addition, for the automatic post-editing task, neural end-to-end systems were found to represent a "significant step forward" over a basic statistical approach. Other recent studies have reported an increase in quality when comparing NMT with SMT using automatic metrics (Bahdanau et al., 2015; Jean et al., 2015) or small-scale human evaluations (Bentivogli et al., 2016; Wu et al., 2016). Wu et al. (2016) report their NMT system outperforming SMT approaches (for English to Spanish, French, simplified Chinese and back), particularly for morphologically rich languages, with impressive human evaluation ratings. Bentivogli et al. (2016) report that English-German NMT post-editing was reduced on average by 26% when

compared with the best-performing SMT system, with fewer word order, lexical, and morphological errors, concluding that NMT has "significantly pushed ahead the state of the art", particularly for morphologically rich languages.

Toral and Sánchez-Cartagena (2017) compare NMT and PBSMT for nine language pairs (English to and from Czech, German, Romanian, Russian, and English to Finnish), with engines trained for the WMT newstest data. Better automatic evaluation results are obtained for NMT output than for PBSMT output for all language pairs other than Russian-English and Romanian-English. NMT systems' increased reordering results in NMT systems performing better than SMT for inflection and reordering errors in all language pairs. However, they also report that SMT appears to perform better than NMT for segments longer than 40 words, when applying the chrF1 automatic evaluation metric (Popović, 2015).

This overview of recent work suggests that NMT has brought great improvement to the field, especially if one considers state-of-the-art automatic evaluation metrics. However, the progress is not always evident. Section 3 presents three use cases in which NMT was compared against SMT and evaluated via human assessments. What emerges is that depending on the different domains and on the various language pairs under study NMT has not always yielded the best results.

3. Use Cases

Each use case focuses on a different domain, and covers a different set of language pairs. First, Section 3.1 looks at NMT for e-commerce, describing important parts of a more extended study that is reported in detail in Calixto et al. (2017b). The second use case (Section 3.2) is an evaluation performed by Iconic Translation Machines Ltd.[1], whose goal was to find out whether NMT could provide better translations for the patent domain than SMT. Finally, the third and last use case (discussed in Section 3.3) is a comparison conducted as part of the EU-funded TraMOOC project on data taken from Massive Open Online Courses (MOOCs) in English.

3.1. NMT for E-Commerce Product Listing

A common use case in e-commerce consists in leveraging MT to make product descriptions, user reviews and comments (e.g. on dedicated forums) as widely accessible as possible, regardless of the customers' native language or country of origin. In previous work, Calixto et al. (2017a) compared the quality of product listings' translations obtained with a multi-modal NMT model against two text-only approaches: a conventional attention-based NMT and a PBSMT model. Translations were evaluated using automatic metrics as well as by means of a qualitative evaluation, whose final goal was to test whether training an NMT system with access to the product images improved the output quality for translations from English into German.

[1] http://iconictranslation.com/

MT Systems - Three different systems were compared in this experiment (1) a *PBSMT* baseline model built with the Moses SMT Toolkit (Koehn et al., 2007), (2) a text-only NMT model (NMT_t), and (3) a multi-modal NMT model (NMT_m), described in more detail in Calixto et al. (2017b), which expands upon the text-only attention-based model and introduces a *visual component* to incorporate *local* visual features.

The data set consists of product listings and images with 23,697 training tuples, each containing *(i)* a product listing in English, *(ii)* a product listing in German, and *(iii)* a product image. Validation and test sets have 480 and 444 tuples, respectively. One point to consider is that the translation of user-generated product listings poses particular challenges, for instance because they are often ungrammatical and can be difficult to interpret even by a native speaker of the language. In particular, the listings in both languages have many scattered keywords and/or phrases glued together, as well as a few typos. These are all complications that make the multi-modal MT of product listings a challenging task, as there are multiple difficulties associated with processing listings and images.

Evaluation - For the qualitative human evaluation, bilingual native German speakers were asked to (1) *assess the multi-modal adequacy* of translations (number of participants N=18); and (2) *rank* translations generated by different models from best to worst (number of participants N = 18). For the *multi-modal adequacy assessment*, participants were presented with an English product listing, a product image and a translation generated by one of the models, without knowing which model. They were then asked how much of the meaning of the source was also expressed in the translation, while taking the product image into consideration, using a 4-point Likert scale (where *4 = None of it* and *1 = All of it*). For the *ranking* assessment, participants were presented with a product image and three translations obtained from different models for a particular English product listing (without identifying the models) and were asked to rank translations from best to worst.

The automatic evaluation was performed with four widely adopted automatic MT metrics: BLEU4, METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and chrF3.

Results - Table 1 contrasts some automatic metrics with human assessments of the adequacy of translations obtained with two text-only baselines, PBSMT and NMT_t, and one multi-modal model NMT_m.

Model	BLEU4↑	METEOR↑	TER↓	chrF3↑	Adequacy↓
NMT_t	22.5	40.0	58.0	56.7	$2.71 \pm .48$
NMT_m	$25.1^{\dagger}$	$42.6^{\dagger}$	$\mathbf{55.5}^{\dagger}$	58.6	$\mathbf{2.36} \pm .47$
PBSMT	$\mathbf{27.4}^{\dagger\ddagger}$	$\mathbf{45.8}^{\dagger\ddagger}$	$\mathbf{55.4}^{\dagger}$	**61.6**	$\mathbf{2.36} \pm .47$

Table 1. Adequacy of translations and four automatic metrics on product listings and images. For the first three metrics, results are significantly better than those of NMT_t (†) or NMT_m (‡) with $p < 0.01$.

The PBSMT model outperforms both the NMT models according to BLEU, METEOR and chrF3. However, there are no differences between the NMT_m model and the PBSMT according to TER scores.

Additionally, the adequacy scores for both these models, NMT_t and PBSMT, are on average the same according to scores computed over human assessments.

Nonetheless, even though both models are found to produce equally adequate output, translations obtained with PBSMT are ranked best by humans over 56.3% of the time, while translations obtained with the multi-modal model NMT_m are ranked best 24.8% of the time. These results suggest that although NMT models can sometimes reach PBSMT automatic MT scores, they are not preferred by human evaluators according to this use-case.

3.2. NMT for the Patent Domain

The evaluation presented in this section was based on a collaborative project carried out between the MT group at the ADAPT Centre, Dublin City University (Ireland), and Iconic Translation Machines Ltd. (Iconic), a commercial MT provider based in Dublin (Ireland). Iconic develops domain-specific MT engines for its users, frequently addressing language pairs and content types that pose great challenges for MT. One such combination in particular demand is Chinese patent information, for translation into English, with more that 100 million words machine translated in 2016.

The goal of this evaluation was to compare the performance between the mature Chinese to English patent MT engines used in production at Iconic with novel NMT engines developed at the ADAPT Centre on an 'apples to apples' basis, trained on the same available data.

The domain of evaluation was chemical patent titles and abstracts (see Table 2). This content type has particular characteristics that present challenges for MT, including very technical content with specialised terminology, names of chemical components, and alphanumeric and aminoacid sequences. The titles and abstract section of the patent themselves are quite distinct: titles are short, with 8.2 tokens on average, and are written in a formulaic telegraphic style; abstracts typically contain between 2-6 sentences that are quite long, with an average length of 42.5 tokens.

Description	Sentence Pairs	Words (source)
Chemical Abstracts	1,076,894	50,198,888
Chemical Titles	350,840	2,868,121
General Patent	11,931,127	324,222,969
Glossaries	1,575	1,575
Total	**13,358,861**	**377,291,553**

Table 2. Training data use for Iconic and NMT engine building

MT Systems - The Iconic MT engines are based on a proprietary Ensemble Architecture ^{TM} which combines elements of phrase-based, syntactic, and rule-driven MT, along with automatic post-editing. The engines have been highly tuned over a number of years for the patent domain, using multiple different translation and language models, and incorporate content-specific terminology.

The ADAPT/Iconic NMT engines were implemented using attention-based models built with Nematus[2] using various combinations of data (given there are slightly different domains, all data is used, i.e. just in-domain data, and in-domain plus different portions of the more general data chosen using data selection). We also tuned on different development sets for titles and abstracts. The four best development engines w for the evaluation. Both engines were trained using the same data, which included a mix of very content-specific in-domain data, more general patent data (including chemistry sub-domain) and technical glossaries.

Evaluation - Engines were evaluated separately on their performance on titles and abstracts, with two different test sets comprising 1,123 segments each. Standard automatic evaluation was carried out, and BLEU scores are reported in Table 3. Human evaluation was also carried out to compare the performance of the two engines. Two reviewers assessed 100 randomly selected segments from the aforementioned test sets in two ways: a blind ranking of the better translation (given a reference), and an error analysis to identify the main translation error in a given segment. The error taxonomy consisted of punctuation, part of speech, omission, addition, wrong terminology, literal translation, and word form. Segments were randomly selected from the test set, so that 25% of the segments were short sentences (i.e. they contained <10 words), 25% were long sentences (i.e. >40 words), and the remaining 50% were medium-length sentences (i.e. between 11 and 39 words).

Automatic evaluation results show that NMT slightly outperformed SMT on titles, whereas the SMT system outperformed NMT on abstracts. Regarding human evaluation, in general the SMT system was ranked 'best' 54% of the times, against 39% for NMT. When looking into sentence length, the SMT system was ranked 'best' 84% of the times for short sentences, against only 8% for the NMT system; and ranked best 58% of the times for long sentences (>40 tokens), against 33% for NMT. The NMT system was ranked 'best' more times than the SMT system only for medium-length sentences (>10<40 words), with 57% of preferences against 36% for SMT.

Results - Error types found in the NMT output were high for omission (37% of errors found in the segments against 8% for the SMT system), whereas for SMT the errors consisted of sentence structure (35% of the segments against 10% for the NMT system).

For segments free of errors, 25% of segments from the SMT system were found not to contain any errors, against only 2% of segments from the NMT system. These results indicate again that the NMT system surpasses the SMT one regarding automatic metrics (for the Titles), but human evaluation still prefers the SMT system.

System	**Titles** (BLEU)	**Abstracts** (BLEU)
Iconic MT	31.99	***28.32***
Neural MT	***37.52***	13.39

Table 3. Automatic MT evaluation results for chemical patent titles and abstracts.

[2] https://github.com/rsennrich/nematus

3.3. NMT for the MOOC domain

The evaluation presented in this section was conducted as part of the EU-funded TraMOOC (Translation for Massive Open Online Courses) project[3], which is a Horizon 2020 collaborative project aiming at providing reliable MT for MOOCs. A PB-SMT and an NMT system were compared across four translation directions (i.e. from English (EN) into German (DE), Greek (EL), Portuguese (PT), and Russian (RU) in a series of extensive assessment tasks. The goal of this comparison was to decide which system would provide better quality translations for the project domain.

MT Systems - The phrase-based SMT used was Moses, and the NMT systems were attentional encoder-decoder networks, which were trained with Nematus. The MT engines were trained on large amounts of training data from various sources: WMT training data[4] and OPUS[5], TED from WIT3[6], QCRI Educational Domain Corpus (QED)[7], a corpus of Coursera MOOCs, and the project's own collection of educational data. The amount of training data used is shown in Table 4.

As this evaluation was intended to identify the best-performing MT system for the translation of MOOCs, test sets were extracted from real MOOC data (one thousand English segments - for the ranking task, just one hundred segments were used). These data included explanatory texts, subtitles from video lectures, or user-generated content (UGC) from student forums or the comment sections of e-learning resources.

The UGC data was often poorly formulated and contained frequent grammatical errors. The other texts presented more standard grammar and syntax, but contained specialized terminology and non-contextual variables and formulae.

Target Language	DE	EL	PT	RU
Out-of-domain	23.78	30.73	31.97	21.30
In-domain	0.27	0.14	0.58	2.31

Table 4. Training data size for training MT engines for EN→ translation direction (number of sentence pairs, in millions).*

Evaluation - For the evaluation, automatic metrics were used (BLEU, METEOR and HTER (Snover et al., 2006)), and human evaluation was also performed. The human evaluation was performed by professional translators (three for EL, PT and RU, and two for DE) and consisted of: i) post-editing (PE) of the MT output to achieve publishable quality in the final revised text, ii) rating of fluency and adequacy (i.e.

[3] http://tramooc.eu/

[4] http://www.statmt.org/wmt16/

[5] http://opus.lingfil.uu.se/

[6] http://www.clg.ox.ac.uk/tedcorpus

[7] http://alt.qcri.org/resources/qedcorpus/

the extent to which a target segment reflects the meaning of the source segment) on a 4-point Likert scale for each segment, and iii) performing error annotation using a simple taxonomy (which included: inflectional morphology, word order, omission, addition, and mistranlsation).

Results - The automatic evaluation (see Table 3) showed that NMT outperformed SMT in terms of BLEU and METEOR scores for German, Greek and Russian (statistically significant in a one-way ANOVA pairwise comparison ($p<.05$)).

For Portuguese, only moderate improvements can be observed. The HTER scores show that more PE was required when using the output from the SMT system for all target languages (not statistically significant). These results indicate that when human intervention was considered (post-editing), the gain with NMT was less consistent.

Lang.	System	BLEU	METEOR	HTER	Fluency	Adequacy
DE	SMT	41.5	33.6	49.0	2.60	2.85
	NMT	61.2 †	42.7 †	32.2	2.95	2.79
EL	SMT	47.0	35.8	45.1	2.86	3.44
	NMT	56.6 †	40.1 †	38.0	3.08	3.46
PT	SMT	57.0	41.6	33.4	3.15	3.73
	NMT	59.9	43.4	31.6	3.22	3.79
RU	SMT	41.9	33.7	44.6	2.70	2.98
	NMT	57.3 †	40.65 †	33.9	3.08	3.12

Table 5. Automatic Evaluation Results (statistically significant results marked with †), Fluency and Adequacy

Human Evaluation - Regarding the human assessment of *fluency*, although no statistically significant differences were found, NMT was rated as more fluent than SMT for all language pairs (Table 5). Results for *adequacy* were less consistent, with higher mean scores for German SMT. These results show that as NMT gains in fluency, however, when assessing how much of the meaning expressed in the source appears in the translation, SMT is slightly better than or equal to NMT.

Regarding the *error annotation* task, the total number of issues identified in the output was greater for SMT than NMT for all language pairs.

Moreover, the number of segments without errors was greater for NMT across all language pairs. NMT output was also found to contain fewer word order errors and fewer inflectional morphology errors in all the target languages. However, SMT output contained fewer errors of omission, addition, or mistranslation for EN-EL than the NMT output; it also showed fewer omissions than the NMT system for EN-PT, while EN-RU SMT showed fewer mistranslations than the NMT system. Interestingly, for German, inflectional morphology errors make up 49% of all

Lang.	System	Technical Effort	Temporal Effort	WPS
DE	SMT	5.8	74.8	0.21
	NMT	3.9	72.8	0.22
EL	SMT	13.9	77.7	0.22
	NMT	12.5	70.4	0.24
PT	SMT	3.8	57.7	0.29
	NMT	3.6	55.19	0.30
RU	SMT	7.5	104.6	0.14
	NMT	7.2	105.6	0.14

Table 6. Technical (keystrokes/segment) and Temporal Post-Editing Effort (secs/segment) and words per second (WPS)

the errors found in NMT output, a higher proportion than for SMT (where inflectional morphology accounts for 43% of the errors). With respect to the *post-editing* tasks, results show that fewer NMT segments were considered by participants to require editing (but with statistical significance only for German ($p<.05$, where $M=.06$, $SE=.04$)). Average throughput or temporal effort (Table 6) was only marginally improved for German, Greek and Portuguese post-editing with NMT, while temporal effort for English-Russian was lower for SMT at the segment level. These results are also replicated in words per second (WPS).

Technical post-editing effort was reduced for NMT in all language pairs using measures of actual keystrokes (Table 6) or the minimum number of edits required to go from pre- to post-edited text (HTER in Table 5). Feedback from the participants indicated that they found NMT errors more difficult to identify, whereas word order errors and disfluencies requiring revision were detected faster in SMT output.

Finally, regarding the *ranking* task, the participants in the evaluation preferred NMT output across all language pairs, with a particularly marked preference for English-German. There was a 53% preference for NMT for short segments (20 tokens or fewer), and a 61% preference for NMT for long segments (over 20 tokens). In conclusion, for the language pairs under consideration (EN-DE, EN-EL, EN-PT and EN-RU) and for the specific MOOC domain, fluency was improved and word order errors decreased when using NMT. Fewer segments required post-editing when using NMT, especially due to the lower number of morphological errors. There was, however, no clear improvement with regard to omission and mistranslation errors when comparing SMT and NMT. There was also no great decrease in post-editing effort, suggesting that NMT for production may not as yet offer more than an incremental improvement in temporal post-editing effort.

4. Discussion and Conclusion

NMT has generated great hype, especially as the translation industry is eager for improved MT quality in order to minimise costs (Moorkens, 2017). Although promising results are being reported when comparing NMT with other MT paradigms using automatic metrics, when human evaluation is added to the comparison, the results are not yet so clear-cut. We have attempted to exemplify this statement with three use-cases comparing NMT against SMT systems where the evaluation was also performed by humans.

The results presented in Section 3.1 for translations of product listings show that NMT models are indeed very promising, especially considering that the state-of-the-art PBMST system has been deployed for quite some time, whereas the NMT models – especially the multimodal NMT system – have been developed over a shorter period of time. However, the PBSMT system still produces better translation when assessed both via automatic and human evaluation metrics. The same outcome can be observed in Section 3.2, with NMT models fast approaching SMT automatic scores

within a few months of deployment for the patent domain. It is important to notice that for both use cases 3.1 and 3.2, the training data is the same training data that is used in their everyday work, which makes it real-world results.

Finally, the extensive human evaluation described in Section 3.3 for the MOOC domain shows that NMT performs well in terms of automatic metrics (apart from Portuguese, where the improvement is only marginal), but is inconsistent for adequacy and post-editing effort. Even though the neural model demonstrates gains in fluency, it also shows a greater number of errors of omission, addition and mistranslation. The decision to move to the NMT model as the MT system of choice for the TraMOOC project reaffirms that neural models are very promising even though little time is put into their development when compared to long-standing PBSMT systems.

While automatic evaluation results published for NMT are undeniably exciting, so far it would appear that NMT has not fully reached the quality of SMT, based on human evaluation. We believe that the hype created in the MT field with the rise of the neural models must be treated cautiously. Overselling a technology that is still in need of more research may cause negativity about MT, as already seen before with SMT systems (especially with the release of the freely-available Moses toolkit in 2006, which made it easier for everyone to train their own MT system), when it was claimed that MT was producing 'near human quality' translations and that MT would 'steal translators' jobs', making translators 'merely post-editors of MT'. The hype that came with this euphoric presentation of SMT systems created a wave of discontent and suspicion among translators, that resulted in an 'us *versus* them' type of confrontation.

NMT no doubt represents a step forward for the MT field. However, there are also limitations for the neural models that cannot be overlooked and still need to be addressed. In our view, at this stage, researchers and industry need to be cautious not to promise too much, and allow for more research to address the limitations of NMT and more extensive human evaluations to be performed, addressing as many text types, domains and language pairs as possible.

Acknowledgements

The TraMOOC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 644333. The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*, San Diego, California., 2015.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. *CoRR*, abs/1608.04631, 2016. URL http://arxiv.org/abs/1608.04631.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

Calixto, Iacer, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. Using Images to Improve Machine-Translating E-Commerce Product Listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 637–643, Valencia, Spain, 2017a. URL http://www.aclweb.org/anthology/E17-2101.

Calixto, Iacer, Qun Liu, and Nick Campbell. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting on Association for Computational Linguistics - Volume 1*, Vancouver, Canada (Paper Accepted), 2017b. URL https://arxiv.org/abs/1702.01287.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. URL http://www.aclweb.org/anthology/D14-1179.

Denkowski, Michael and Alon Lavie. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, 2014. URL http://www.aclweb.org/anthology/W14-3348.

Jean, Sébastien, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, 2015. URL http://www.aclweb.org/anthology/P15-1001.

Kalchbrenner, Nal and Phil Blunsom. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1700–1709, Seattle, October 2013.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Luong, Thang, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 1412–1421, Lisbon, Portugal, 2015. ISBN 978-1-941643-32-7.

Moorkens, Joss. Under pressure: translation in times of austerity. *Perspectives*, 25(3):1–14, 2017. doi: 10.1080/0907676X.2017.1285331. URL http://dx.doi.org/10.1080/0907676X.2017.1285331.

Popović, Maja. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015.

Sennrich, Rico and Barry Haddow. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany, August 2016.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany, 2016.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200(6), 2006.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, Montréal, Canada, 2014.

Toral, Antonio and Víctor M. Sánchez-Cartagena. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-1100.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144, 2016.

Address for correspondence:
Sheila Castilho
sheila.castilho@adaptcentre.ie
ADAPT Centre, Dublin City University